Microsoft Research

Microsoft

Beyond Open Access to Open Data

Tony Hey
Vice President
Microsoft Research

Big Data and the Fourth Paradigm

Data Size and Speed are Growing

Entire sequence of DNA for the human body, consists of around 3 billion of these base pairs.

The human genome requires ~750 megabytes of storage

Large Hadron Collider

150 million sensors delivering data 40 million times per second.

Data flow: ~700 MB/sec

~15 PB/year

1000's of scientists around the world; Institutions in 34 different countries:

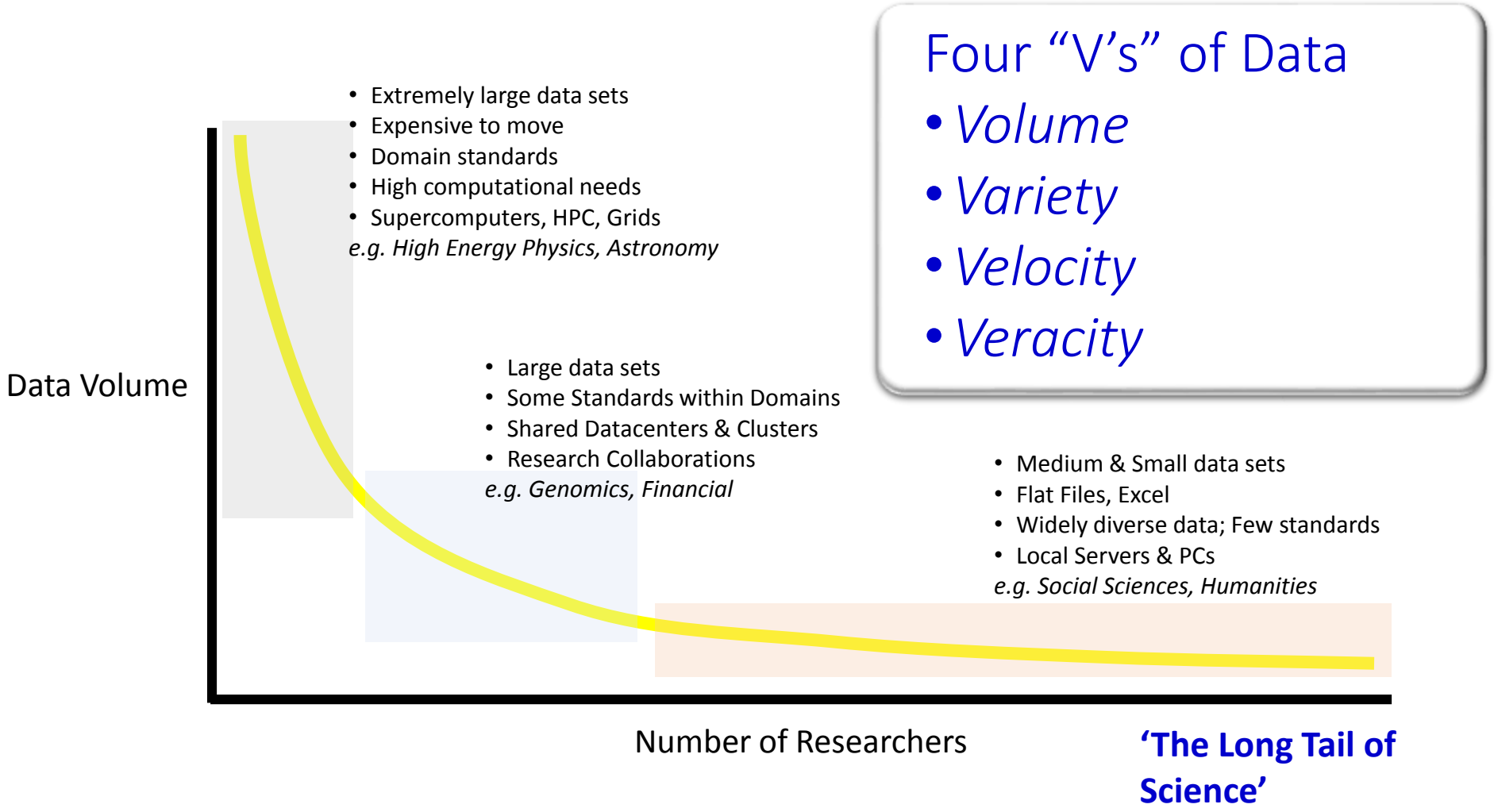
Thousands of small antennas spread over a distance of more than 3000km.

Data flow: ~60 GB/sec

1 Million PB/day

The SKA supercomputer will perform 1018 operations per second ~ 100M PCs

Much of Science is now Data-Intensive



Jim Gray, Turing Award Winner

The ‘Cosmic Genome’ Project

- The Sloan Digital Sky Survey was the first major astronomical survey project:
  - 5 color images and spectra of ¼ of the sky
  - Pictures of over 300 million celestial objects
  - Distances to the closest 1 million galaxies

- Jim Gray from Microsoft Research worked with astronomer Alex Szalay to build the public ‘SkyServer’ archive for the survey

➢ New model of scientific publishing - publish the data <u>before</u> astronomers publish their analysis

eScience and the Fourth Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

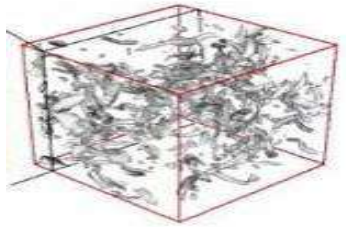
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

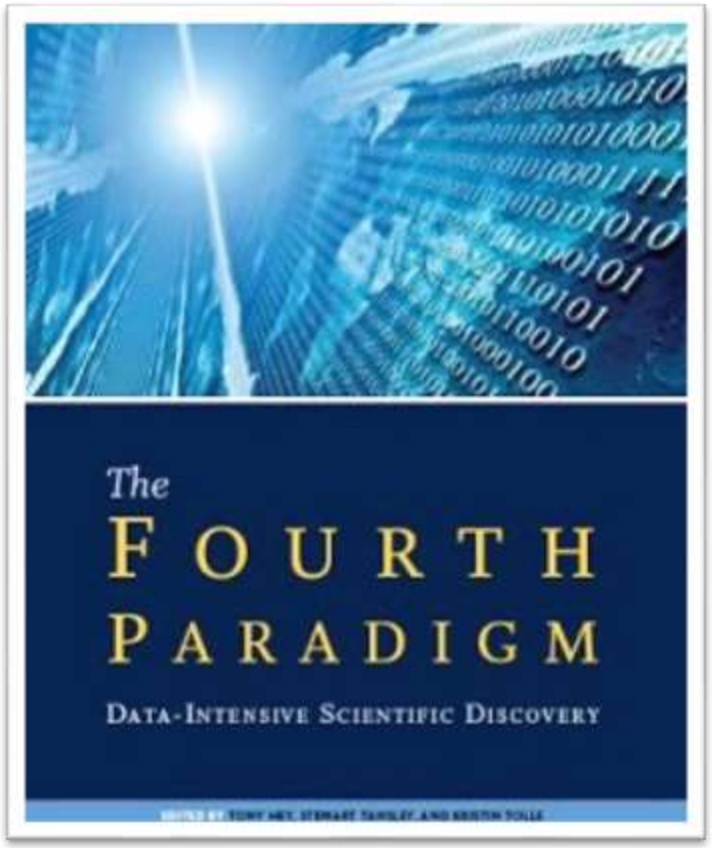


eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

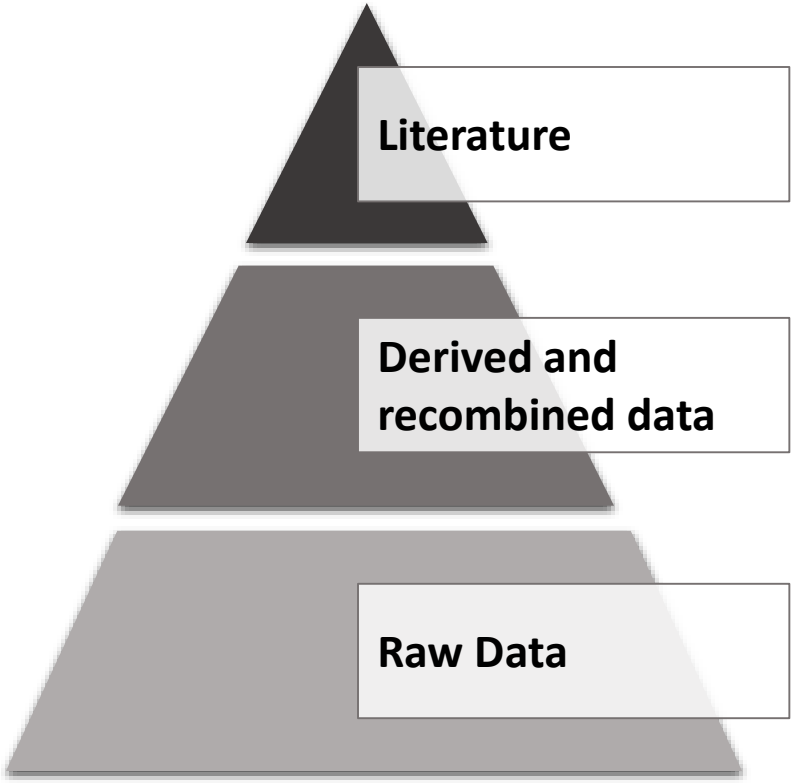
(With thanks to Jim Gray)

eScience and Data-Intensive Scientific Discovery



All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips –
  For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



(*From Jim Gray's last talk)*

Open Access: The Tipping Point

US Fair Access to Science and Technology Research (FASTR) Act

- *"The United States has a substantial interest in maximizing the impact of the research it funds by enabling a wide range of reuses of the peer-reviewed literature reporting the results of such research, including by enabling automated analysis by state-of-the-art technologies."*
- Federal agencies consider whether or not the terms of use should include *"a royalty free copyright license that is available to the public and that permits the reuse of those research papers, on the condition that attribution is given to the author or authors of the research and any others designated by the copyright owner"*

13 February 2013

US White House Memorandum

- Directive requiring the major Federal Funding agencies *"to develop a plan to support increased public access to the results of research funded by the Federal Government."*
- The memorandum defines digital data *"as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens."*

22 February 2013

Global Research Council

http://www.globalresearchcouncil.org/

- Newly founded network of national research funders from all over the world has endorsed an Action Plan towards Open Access.
- Action Plan consists of the 14 points including
  - Collect and document best practices for rewarding the provision of open access
  - Work with scholarly societies to transition society journals into open access
  - Work with repository organisations to develop efficient mechanisms for harvesting and ac-cessing information

30 May 2013

G8 Science Ministers

- "In a joint statement proposing “new areas” of scientific collaboration for the countries, the ministers say they “recognise the potential benefits of immediate global access to and unrestricted use of published peer-reviewed, publicly funded research results.”
- “We share the intention, therefore, to continue our cooperative efforts and will consider how best to address the global promotion of increasing public access to the results of publicly funded published research including to peer-reviewed published research and research data.”

12 June 2013

University of California approves Open Access

- UC is the largest public research university in the world and its faculty members receive roughly 8% of all research funding in the U.S.
- UC produces 40,000 publications per annum corresponding to about 2 – 3 % of all peer-reviewed articles in world each year
- The faculty remains committed to working with publishers to transform the publishing landscape in ways that are sustainable and beneficial to both the University and the public.

2 August 2013

Stevan Harnad and Green Open Access

- The "Subversive Proposal" was an Internet posting by Stevan Harnad on June 27 1994

- His proposal called on all authors of "esoteric" writings—written only for research impact, not for royalty income—to archive them free online

- This became ‘Green Open Access’ or self-archiving

Dec 2001 – Budapest Open Access Initiative

- The Budapest Open Access Initiative arose from a small but lively meeting convened in Budapest by the Open Society Institute (OSI) on December 1-2, 2001.
- The purpose of the meeting was to accelerate progress in the international effort to make research articles in all academic fields freely available on the internet.

US NIH Open Access Policy

- *Once posted to PubMed Central, results of NIH-funded research become more prominent, integrated and accessible, making it easier for all scientists to pursue NIH's research priority areas competitively.*
- *PubMed Central materials are integrated with large NIH research data bases such as Genbank and PubChem, which helps accelerate scientific discovery.*
- *The Policy allows NIH to monitor, mine, and develop its portfolio of taxpayer funded research more effectively, and archive its results "in perpetuity"*

NIH Open Access Compliance

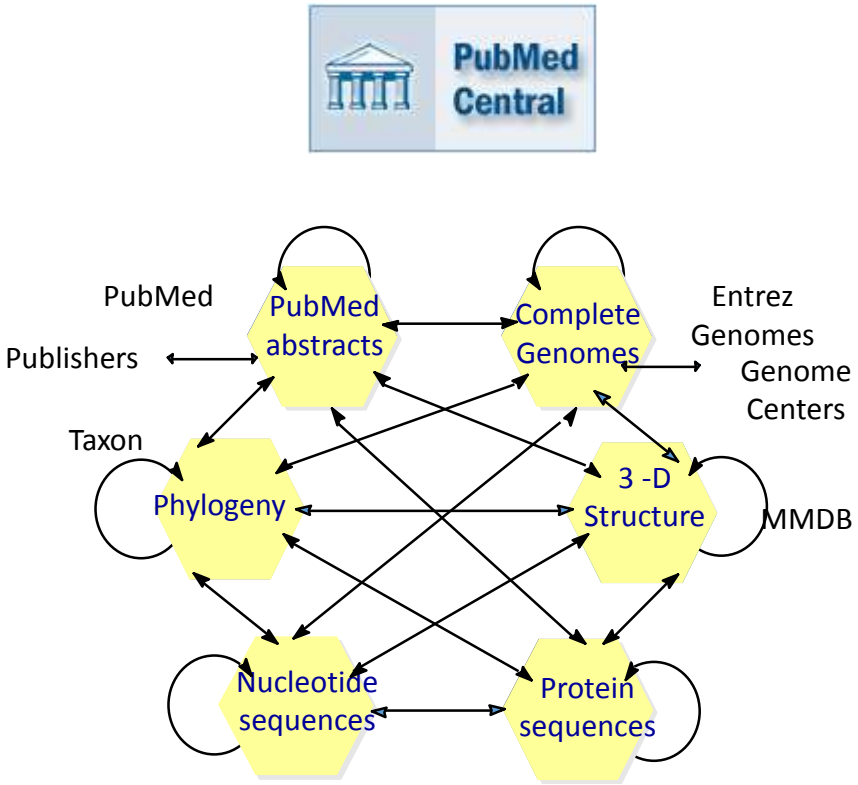
- PMC Compliance Rate
  - Before legal mandate compliance was 19%
  - Signed into law by George W. Bush in 2007
  - After legal mandate compliance up to 75%

- NIH have taken a further step of announcing that, 'sometime in 2013' they

  '… will hold processing of non-competing continuation awards if publications arising from grant awards are not in compliance with the Public Access Policy.'

The US National Library of Medicine

- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

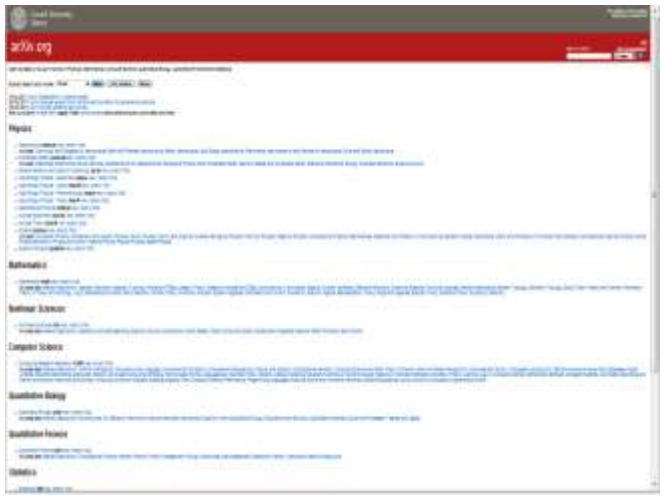


Entrez cross-database search

Green Open Access:
An Existence Proof

Paul Ginsparg and arXiv

- The particle physics research community had a long tradition – since the 1960's - of circulating 'preprints' prior to peer review
- With the advent of word processing and the Web, Paul Ginsparg set up the LANL preprint archive at xxx.lanl.gov in 1991
- The repository expanded to other disciplines and changed its name to arXiv.org in 1999



Comments (1)

- arXiv is over 20 years old and has 7,000 e-prints submitted per month
- Full text versions of over half a million research papers are available free
- More than 200,000 articles downloaded each week by about 400,000 users.
- The arXiv repository ‘publishes’ e-prints prior to the refereeing process
- Acceptance in a journal is effectively a post-publication quality mark

Comments (2)

- Having multiple, slightly different versions of a paper is not a serious drawback in practice. See Gentil-Beccot, Mele and Brooks:

  ‘Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories’.

- This shows a significant citation advantage for papers first posted in arXiv and subsequently published in journals. The paper is, of course, available as arXiv:0906.5418.

Comments (3)

'arXiv is the primary daily information source for hundreds of thousands of researchers in physics and related fields. Its users include 53 physics Nobel laureates, 31 Fields medalists and 55 MacArthur fellows, as well as people in countries with limited access to scientific materials. The famously reclusive Russian mathematician Grigori Perelman posted the proof for the 100-year-old Poincaré Conjecture solely in arXiv.'

http://phys.org/news142785151.htm

The American Physical Society and arXiv?

"Marty Blume, when he was editor-in-chief in the '90's, was incredibly supportive of arXiv. They decided that the membership wanted it, so they'd figure a way to work with it, one way or another. It made me proud to be a physicist."

Paul Ginsparg

The arXiv Sustainability Model

- Operation of arXiv is currently funded by Cornell University Library.
- In 2010, Cornell broadened funding support for arXiv by asking institutions to make an annual contribution based on the amount downloaded by each institution.
- Annual donations vary in size between $2,300 to $4,000, based on usage.
- As of February 2010, 27 institutions have pledged support on this basis.
- The annual budget for arXiv was $400,000 for 2010.

Open Data and Open Science

The Berlin Declaration 2003

- ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’
- Defines open access contributions as including:

  ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’

Scholarship is changing …

- Funding Agencies are now beginning to require curation and preservation of research data
- There is an increased need for research reproducibility
- Governments and funding agencies looking towards data sharing and interoperability
- Need to change career and reward system to value dataset curation and management

Datacite and ORCID

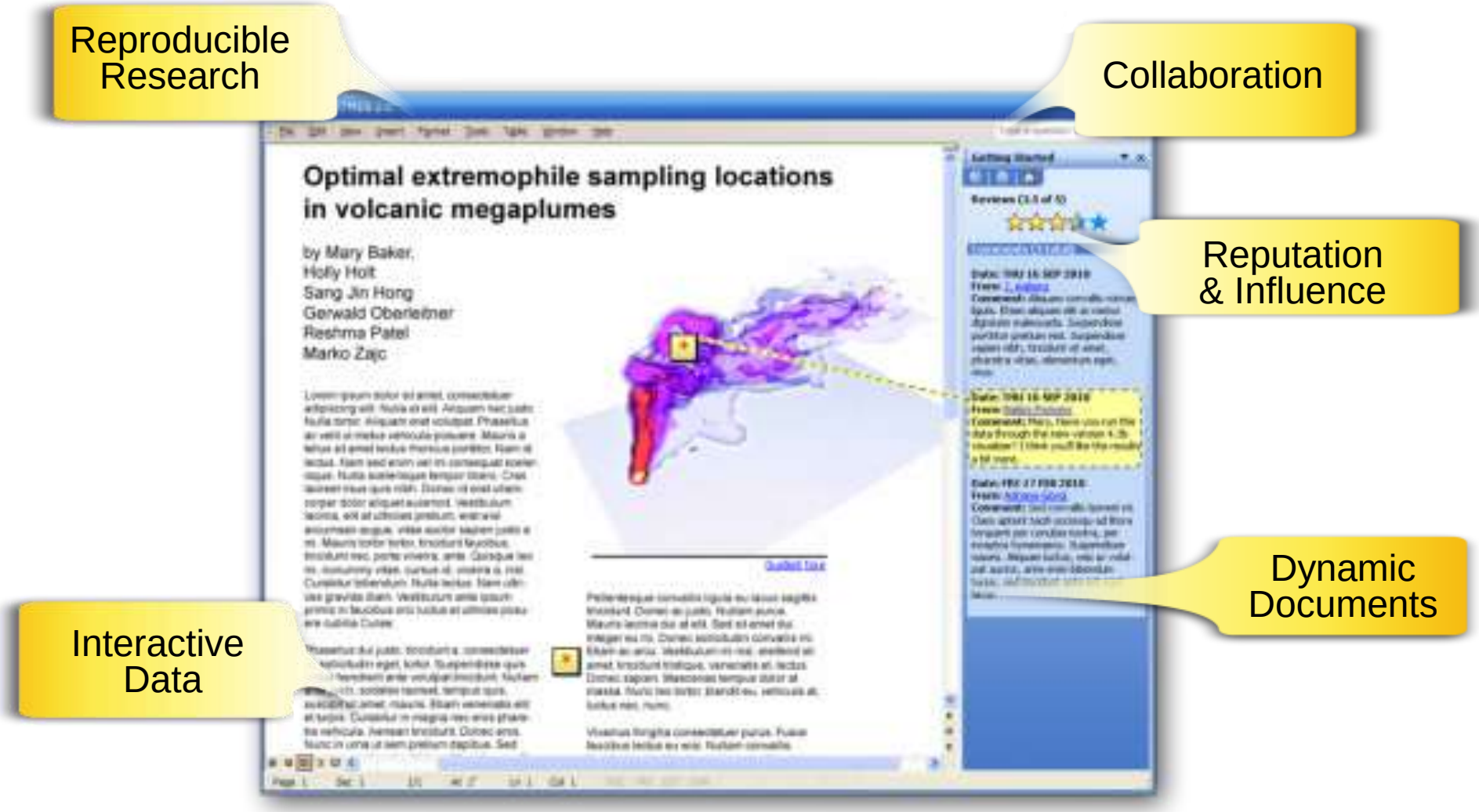
DataCite

- International consortium to establish easier access to scientific research data
- Increase acceptance of research data as legitimate, citable contributions to the scientific record
- Support data archiving that will permit results to be verified and re-purposed for future study.

ORCID - Open Research & Contributor ID

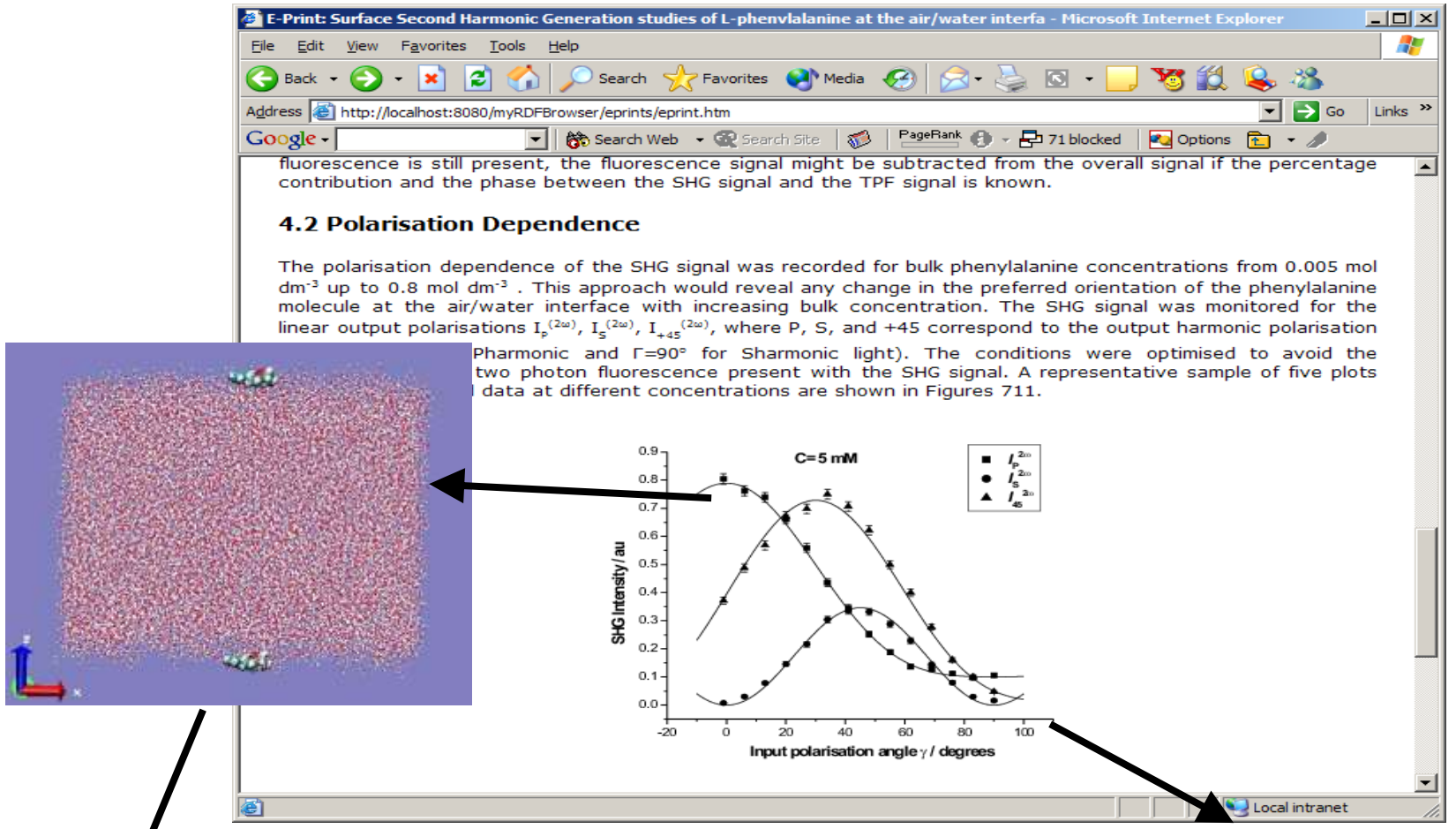
- Aims to solve the author/contributor name ambiguity problem in scholarly communications
- Central registry of unique identifiers for individual researchers
- Open and transparent linking mechanism between ORCID and other current author ID schemes.
- Identifiers can be linked to the researcher’s output to enhance the scientific discovery process

Vision for a New Era of Research Reporting



(*Thanks to Bill Gates SC05*)

Publications as Live Documents



Link to simulation software and data in archive

Link to data, follow links back to the raw data archive

Collaboration and Sharing of Data is Expected and Growing

… expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.

NIH reaffirms its support for the concept of data sharing. We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health … The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.

A primary goal of Data.gov is to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications). Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government.

NSF Data Sharing Policy 2010

"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing."

All future grant proposals now require a two-page Data Management Plan that addresses the above requirement and the Plan will be subject to peer review.

Key driver from a UK Research Council

EPSRC Policy Framework on research data (May 2011)

- "all institutions in receipt of their funding should develop a clear roadmap for research data management, which should be implemented by May 1st 2015"
- "organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years"

PLOS' New Data Policy: Public Access to Data

By Liz Silva
Posted: February 24, 2014

UPDATE: A flurry of interest has arisen around the revised PLOS data policy that we announced in December and which will come into effect for research papers submitted next month. We are gratified to see a huge swell of support for the ideas behind the policy, but we note some concerns about how it will be implemented and how it will affect those preparing articles for publication in PLOS journals. We'd therefore like to clarify a few points that have arisen and once again encourage those with concerns to check the details of the policy or our FAQs, and to contact us with concerns if we have not covered them.

Is the policy about what to share, or about how and where to share it?

There is nothing new in the policy about what types and forms of data should be shared. As we said in December, "PLOS journals have requested data be available since their inception, but we believe that providing more specific instructions for authors regarding appropriate data deposition options, and providing more information in the published article as to how to access data, is important for readers and users of the research we publish." As we have further clarified, "the Data Policy states the 'minimal dataset' consists "of the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety. This does not mean that authors must submit all data collected as part of the research, but that they must provide the data that are relevant to the specific analysis presented in the paper." The 'minimal dataset'

Linking Publications to Data: The State of the Art

Astrophysics Data System ADS

· Find Similar Abstracts (with default settings below)

Toggle Highlighting

· Custom Format
· Electronic Refereed Journal Article (HTML)
· Full Refereed Journal Article (PDF/Postscript)
· FIND IT @ HARVARD
· arXiv e-print (arXiv:astro-ph/0412451)
· On-line Data
· References in the article
· Citations to the Article (84) (Citation History)
· Refereed Citations to the Article
· SIMBAD Objects (3)
· NED Objects (1)
· Also-Read Articles (Reads History)
·
· Translate This Page

Links to e-resources

Links to data

Links to objects

Title: Bow Shock and Radio Halo in the Merging Cluster A520

Authors: Markevitch, M.; Govoni, F.; Brunetti, G.; Jerius, D.

Affiliation: AA(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138; Space Research Institute, Russian Academy of Sciences, 84/32 Profsoyuznaya Street, Moscow 117997, Russia. maxim@head.cfa.harvard.edu), AB(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AC(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AD(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138 maxim@head.cfa.harvard.edu)

Publication: The Astrophysical Journal, Volume 627, Issue 2, pp. 733-738. (ApJ Homepage)

Publication Date: 07/2005

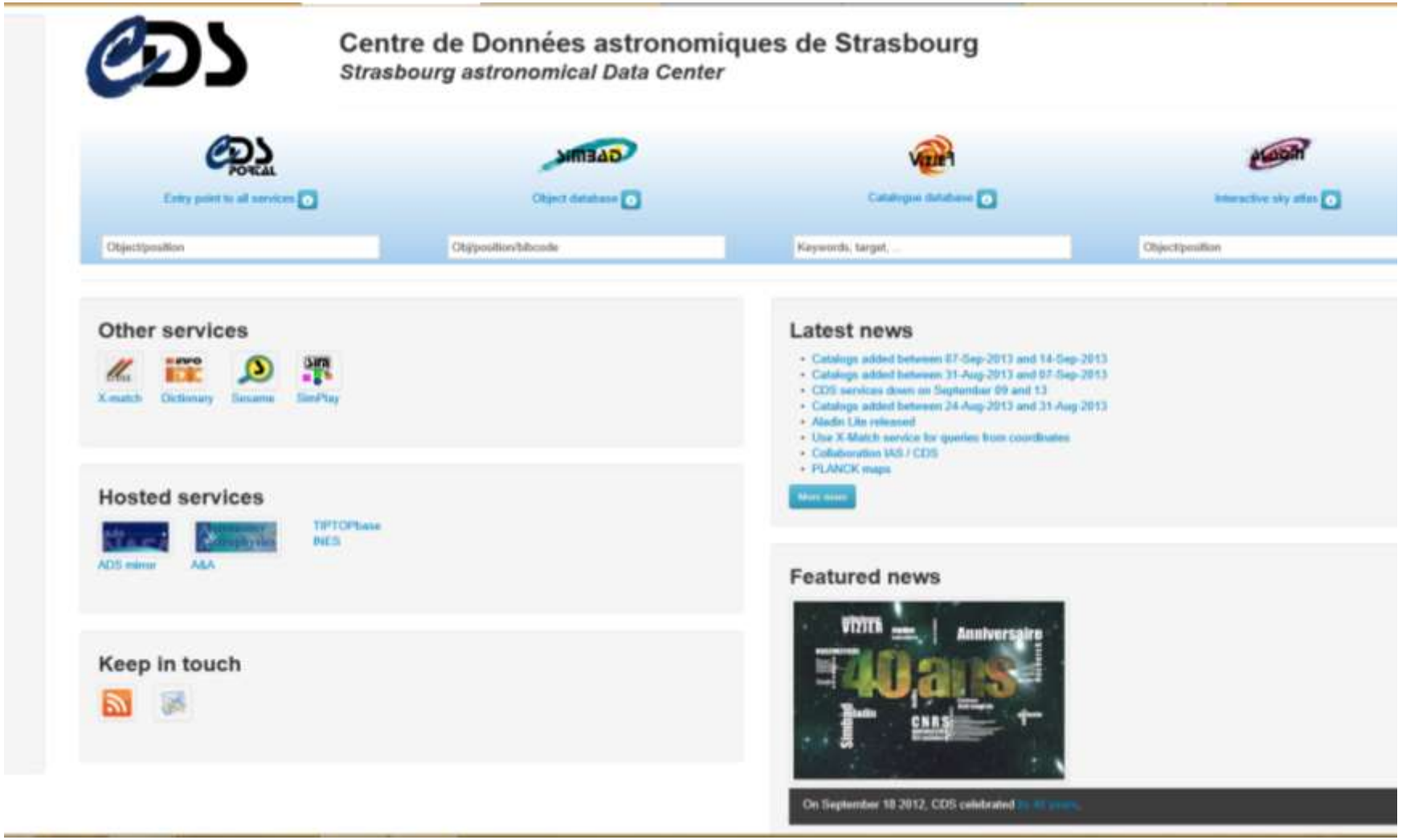
Origin: UCP

Astronomy Keywords: Galaxies: Clusters: Individual: Alphanumeric: A520, Galaxies: Intergalactic Medium, Radio Continuum: General, X-Rays: Galaxies: Clusters

DOI: 10.1086/430695

Bibliographic Code: 2005ApJ...627..733M

Strasbourg CDS Datasets



Literature

WIKIPEDIA
The Free Encyclopedia

AstroBetter

Blogs, Wikis, etc.

"Seamless Astronomy" (Tools)

ads labs

Astrometry.net

flickr

WorldWide Telescope

TOPCAT

ds9

Data

VizieR

SIMBAD

Astro Grid

VAO

IVOA

EURO-VO

AIDA Astronomical Infrastructure for Data Access

"Registries"

DataScope

Disclaimer: *This slide shows key excerpts from within the astronomy community & excludes more general s/w that is used, such as Papers, Zotero, Mendeley, EndNote, graphing & statistics packages, data handling software, search engines, etc.*

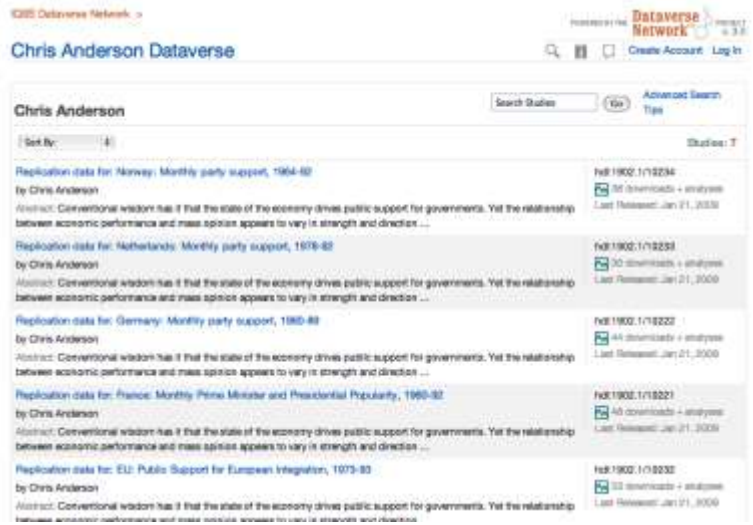
Reinforcing the Link between Research Publications and Research Data

The Dataverse project at Harvard has been awarded an Alfred Sloan Foundation grant for the next 2 years to enhance the link between journals and data

Open Journal System

Seamless integration between the two systems:

Deposit Data to Dataverse through standard API (based on SWORD)



The Dataverse Network

Sustainability of Data Links?

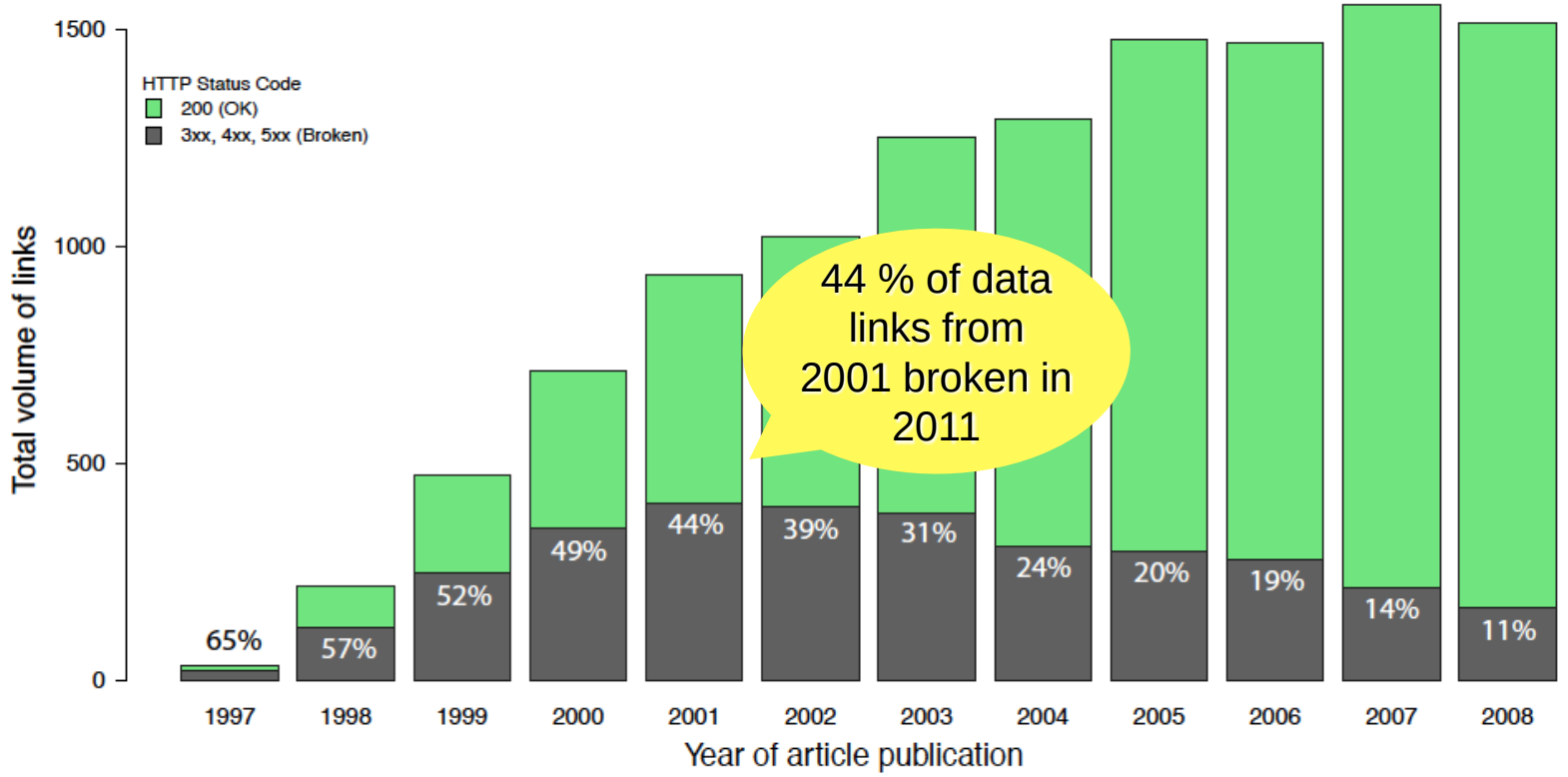
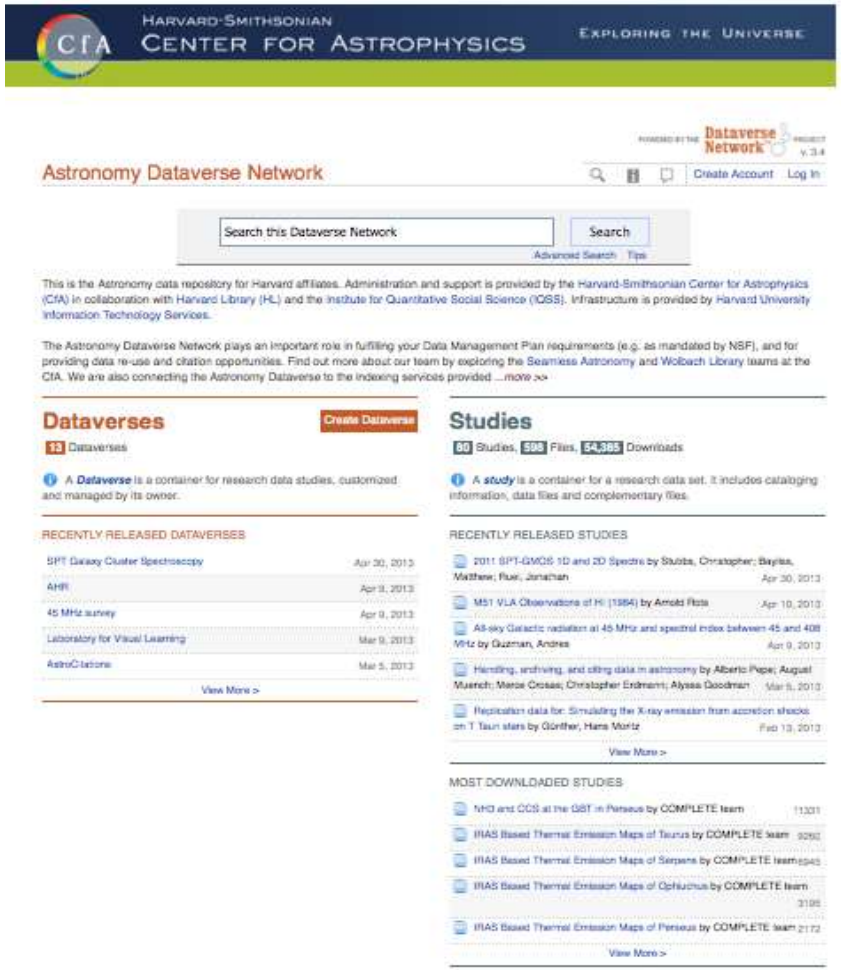


Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

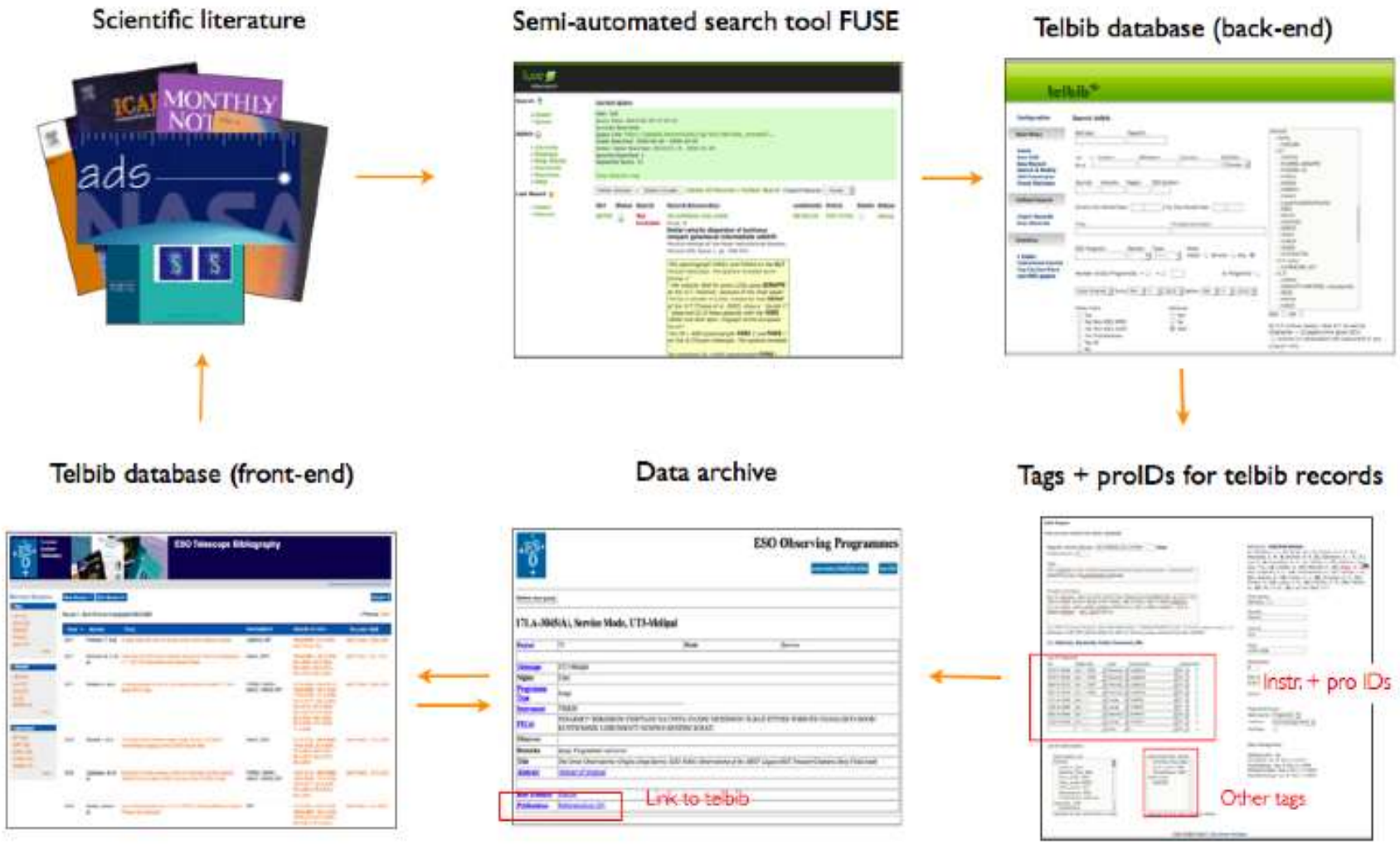
Pepe et al. 2012

Linking to All Data, Big and Small



Slide courtesy of Christopher Erdman

AstroCurator: Telescope Bibliographies

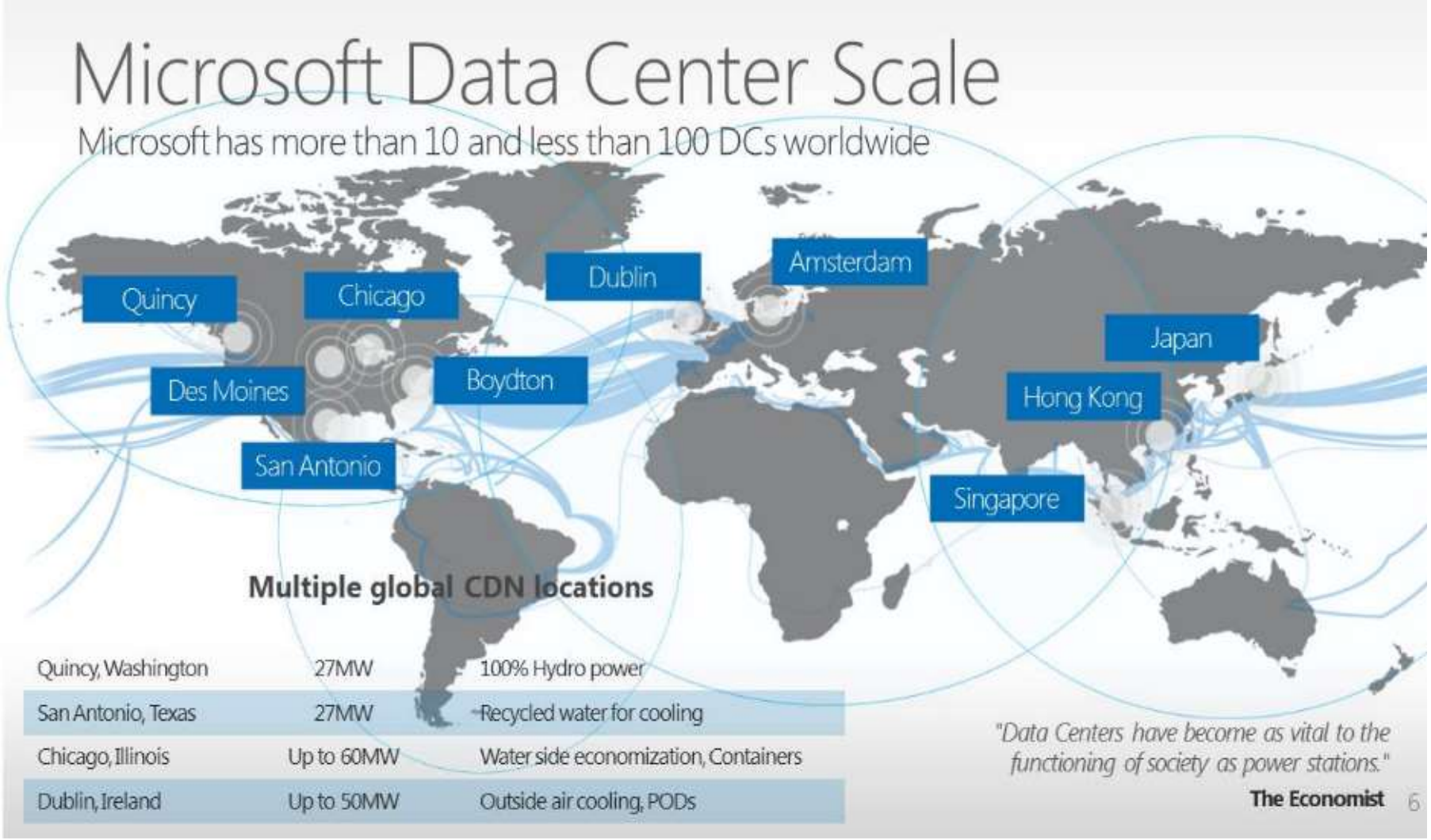


Meakins & Grothkopf, 2011: Linking Publications and Observations: The ESO Telescope Bibliography. ASP Conf. Proc. 461, 767 http://adsabs.harvard.edu/abs/2012ASPC..461..767M

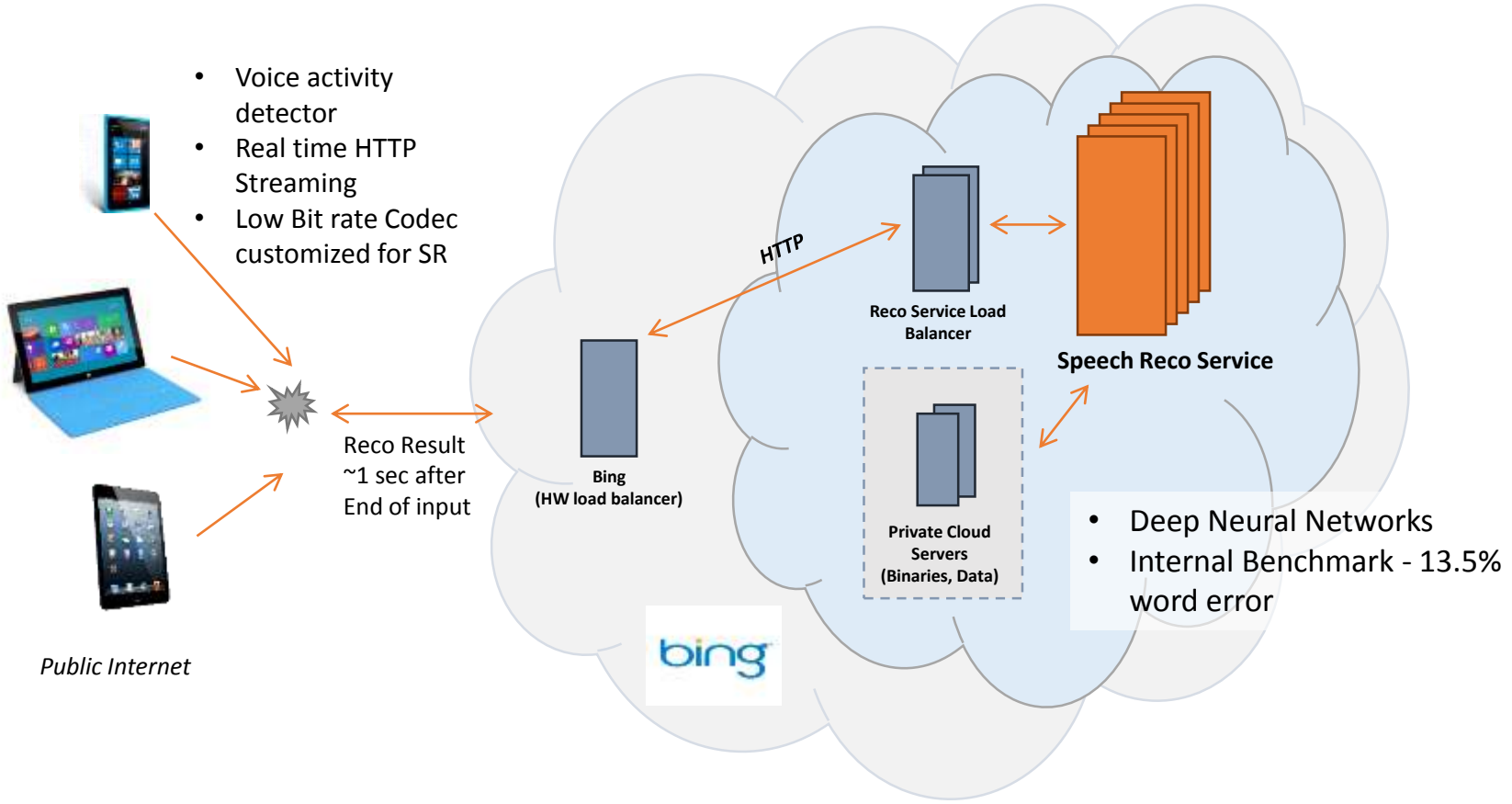
Slide courtesy of Christopher Erdman

Data Science in the Future?

Industry is building out massive Cloud Infrastructure



Bing Speech Recognition Service: The Cloud Changes the Game



Data Repositories

Software

- Traditional Repository Software
  - DSpace (DuraSpace)
  - EPrints (Southampton)
- Data-specific
  - CKAN (OKFN)
  - DataVerse (Harvard)

Hosted

- Dryad (UNC Chapel Hill)
- figshare (Digital Science)

Hybrid approach

- Duracloud (using cloud for preservation/storage) but repository still run locally

Other approaches

- Host CKAN on Azure
- Eprints as a Cloud Service
- DataVerse on Azure

Windows Azure for Research

Accelerate the Speed of Scientific Discovery

Windows Azure provides researchers with the power and scalability of cloud computing for collaboration, computation, and data-intensive processing. This open and flexible global cloud platform supports any language, tool, or framework.

The Windows Azure for Research program:

- Free access to Windows Azure cloud computing and storage (submit proposals for Windows Azure Research Awards)
- Windows Azure for Research training classes
- Support and technical resources

Apply the power of cloud computing to your computational and data challenges. Experiment at **azure4research.com.**

CERN's ZENODO

Zenodo: A new grey literature and data publication solution from CERN

Written by Chris Erdmann
September 12th, 2013

Galactic Gazette
Subscribe RSS

I've been with the Harvard-Smithsonian Center for Astrophysics as the Head Librarian for 3+ years, but a patron request that the John G. Wolbach Library received during my first few weeks still clings to the back of mind. It involved a graduate student simply wishing to submit her dissertation to the Library in electronic format. At the time, we had no solution to manage and disseminate her dissertation as part of an online collection, so we ultimately took the PDF and placed it on our shared network drive. It drove me crazy that we didn't have a solution for both preservation and dissemination of the dissertation. I've continuously revisited the problem, always scanning for potential solutions, only to find that they fell short in some way. That all changed though when I recently started working with the talented development team at CERN behind Zenodo.org, led by Tim Smith and Lars Holm Nielsen.

Image Credit:
Image Above, The Large Hadron Collider/ATLAS at CERN, CC Image Courtesy of Image Editor on Flickr

Posted in Wolbach Library

No Comments

SCIENTIFIC DATA

Helping you publish, discover, and reuse research data

Calling for submissions in Fall 2013, launching in Spring 2014
nature.com/scientificdata

nature publishing group npg

Arkivum

Every bit archived

Arkivum

- Archive as a Service »
- Partners »
- Resources »
- News & Events
- About Us
- Blog

Arkivum provides a large scale, long term, and cost effective digital archiving service with a unique 100% data integrity guarantee.

- ✓ Reduce Capital Expenditure.
- ✓ Free up time for your IT team.
- ✓ Sleep at night knowing that your data is safe.

Message of the Day

New webinar: 13 March with our partner S3. Does your company archive valuable data? Discover the benefits of long-term archiving Register here

Data Repository Registries

Purdue University
603 repositories
http://databib.org

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

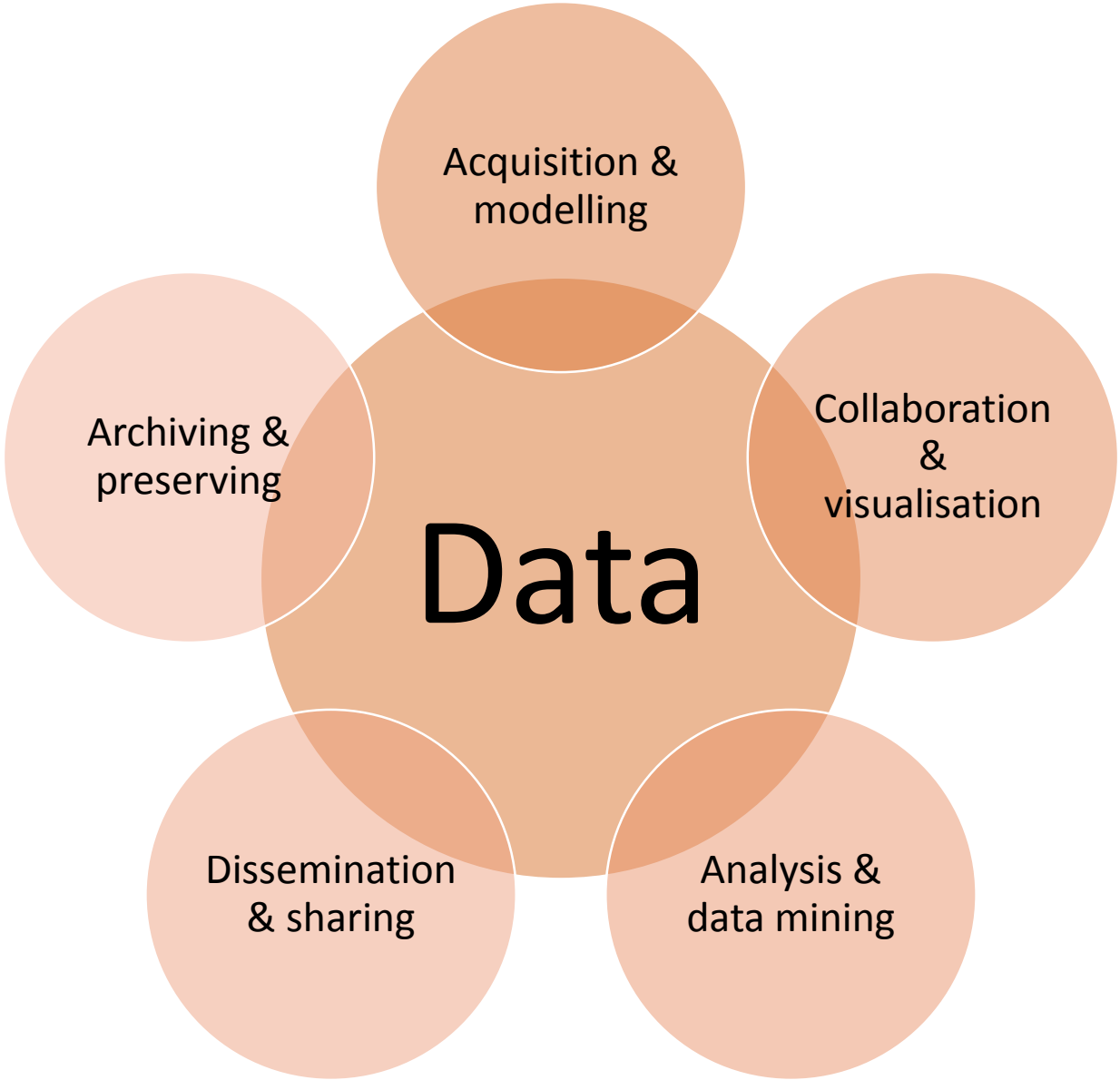
Humboldt-Universität zu Berlin
603 repositories
http://re3data.org

Identify and locate online repositories of research data

- What repositories are appropriate for a researcher to submit his or her data to?
- How do users find appropriate data repositories and discover datasets that meet their needs?
- How can librarians help patrons locate and integrate data into their research or learning?

The goal of re3data.org is to create a global registry of research data repositories. The registry will cover research data repositories from different academic disciplines. re3data.org will present repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions. In the course of this mission re3data.org aims to promote a culture of sharing, increased access and better visibility of research data.

Supporting the Entire Data Life Cycle



254,000 RESULTS

The **Data Scientist** role is a role of the future!
www.**datascientists**.net
The **Data Scientist** role is a role of the future! Future proof your career and start transitioning today.

Data Scientist: The Hottest Job You Haven't Heard Of - Careers ...
jobs.aol.com/articles/2011/08/10/**data-scientist**-the-hottest-job...
Aug 10, 2011 · Data **scientists** are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities

LinkedIn's Monica Rogati On "What Is A **Data Scientist**?" - Forbes
www.forbes.com/.../linkedins-monica-rogati-on-what-is-a-**data-scientist**
Nov 27, 2011 · To continue our series on the emerging role of the **data scientist** in today's data-driven organizations, we spoke with Monica Rogati, Senior Data ...

Related searches for **"data scientist"**

Data Scientist **Seattle**
Data Scientist **Salary**
Data Scientist **Interview Ques...**
Data Scientist **Fortune**
Data Scientist **Jobs**
Introduction to Data **Science**

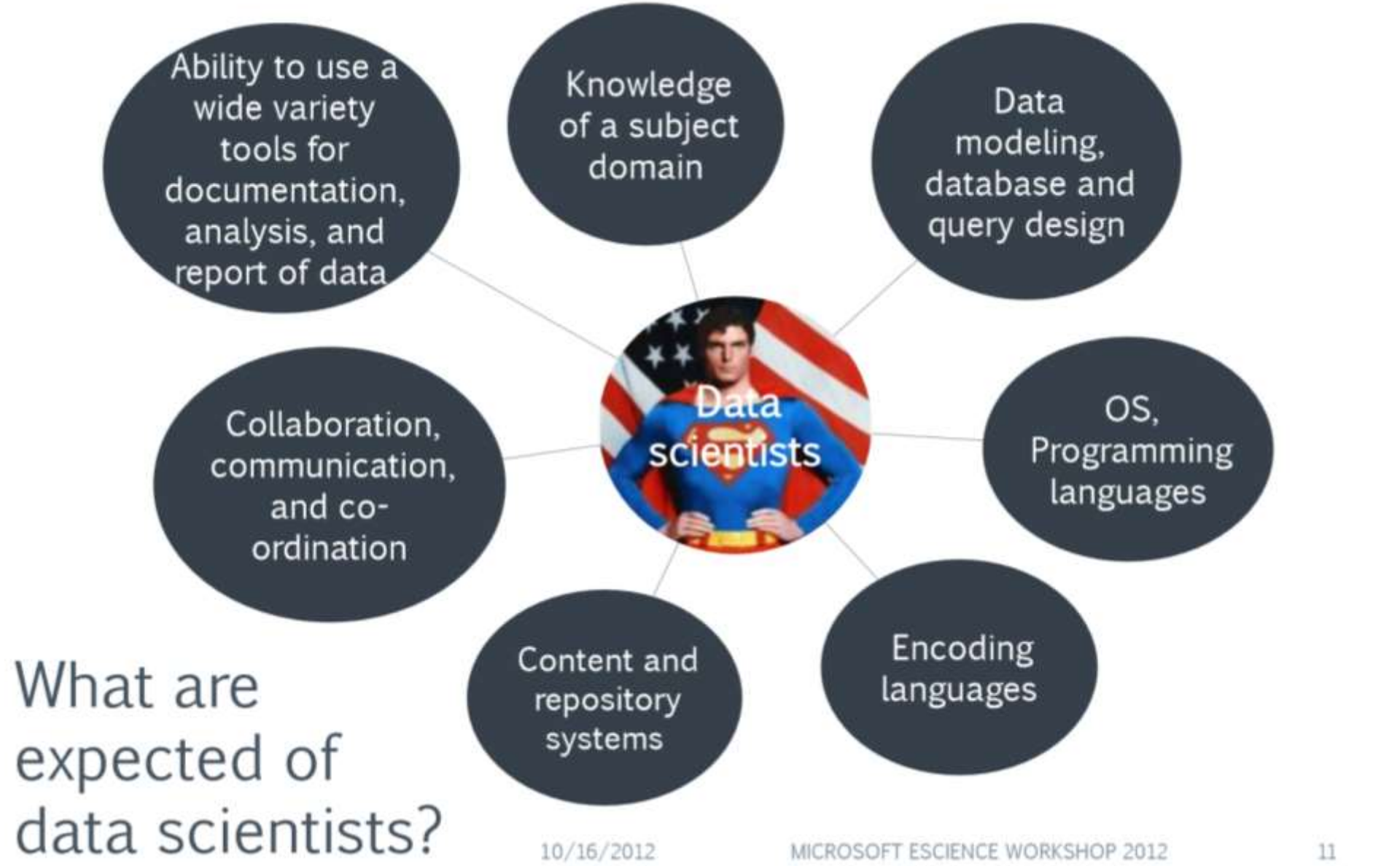
Data scientist: The hot new gig in tech - Fortune Tech
tech.fortune.cnn.com/2011/09/06/**data-scientist**-the-hot-new-gig-in-tech
Sep 06, 2011 · Companies that want to make sense of all their bits and bytes are hiring so-called data **scientists** - if they can find any. FORTUNE -- The unemployment rate ...

The **Data Scientist** | Mine, Visualize, and Learn
www.**thedatascientist**.com
As I jumped from room to room on Turntable.fm last night my eyes caught a glimpse of a rare room titled "AOKIxSOLREPUBLIC" . I clicked it with a fury.

What is a Data Scientist?

Data Engineer 	**People who are expert at** • Operating at low levels close to the data, write code that manipulates • They may have some machine learning background. • Large companies may have teams of them in-house or they may look to third party specialists to do the work.
Data Analyst 	**People who explore data through statistical and analytical methods** • They may know programming; May be an spreadsheet wizard. • Either way, they can build models based on low-level data. • They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.
Data Steward 	**People who think to managing, curating, and preserving data.** • They are information specialists, archivists, librarians and compliance officers. • This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

What is a data scientist? Microsoft UK Enterprise Insights Blog, Kenji Takeda
http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/01/31/what-is-a-data-scientist.aspx



Slide courtesy of Jian Qin

Some Resources

- Microsoft Research
  - http://research.microsoft.com
- Microsoft Research Connections
  - http://research.microsoft.com/en-us/collaboration/
- Science at Microsoft
  - http://www.microsoft.com/science
- Scholarly Communications
  - http://www.microsoft.com/scholarlycomm
- Azure Cloud for Research
  - http://research.microsoft.com/en-us/projects/azure/default.aspx
- Outercurve Foundation
  - http://www.outercurve.org/
- Tony Hey on eScience
  - http://tonyhey.net/